Graph theoretical elaboration of cumulative scaling techniques

Stokman, Frans N.

Veröffentlichungsversion / Published Version
Sammelwerksbeitrag / collection article

Frans N. Stokman

Graph Theoretical Elaboration of Cumulative Scaling Techniques

In this paper a graph theoretical elaboration of the *stochastic* cumulative scaling model of Mokken[1] will be given to determine:

a. cumulative scales of *vertices* on the basis of their relations with other vertices in a simple graph;
b. cumulative scales of relations in a multigraph.

In Stokman both graph theoretical elaborations are given and applied to (co) sponsorship of resolutions in the United Nations General Assembly to determine leadership structures among developing nations[2]. Felling elaborated the applicability of the *deterministic* cumulative scaling model of Guttman to determine cumulative scales of relations in multigraphs[3]. Graph theoretical elaboration of the stochastic scaling model gives however certain new insights, that were not considered by Felling.

In Section 1 we introduce a number of graph theoretical concepts that will be used in the remainder of the paper. In Section 2 cumulative scales of vertices in a simple graph are treated, in Section 3 cumulative scales of relations in a multigraph. After determination of the different cumulative dimensions of relations in a multigraph, for each dimension a new graph can be generated and analyzed. In Section 4 it will be shown how this can be done.

1. Graph Theoretical Concepts[4]

A *graph* is an object, that contains vertices and edges, each edge being incident with one or two vertices[5]. Let us consider , as an example, the graph that consists of the vertices u, v, w, x, y, z and the edges a, b, c, d, e, f, g, h. The incidence relations are given in Table 1. The graph is given in Figure 1.

[1] Mokken, R. J., A Theory and Procedure of Scale Analysis. With applications in political research, The Hague 1970.

[2] See Stokman, Frans N., Roll Calls and Sponsorship. A methodological analysis of Third World group formation in the United Nations, Leiden 1977.

[3] Felling, A. J. A., Sociaal-netwerkanalyse, Alphen aan den Rijn 1974.

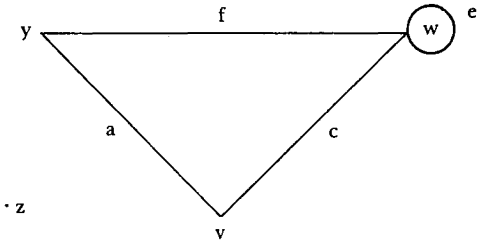
[4] The presentation of the graph theoretical concepts is strongly based on Helmers, H. M., et.al.

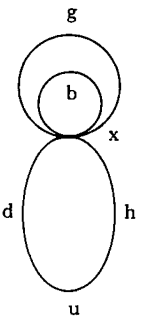
Table 1: Incidence Relations

edge	*incident with*
a	v and y
b	x
c	w and v
d	u and x
e	w
f	y and w
g	x
h	u and x

Source: Helmers et. al., 1975, p. 112.

Figure 1: A Graph





Source: Helmers et. al., 1975, p. 115.

The edges e, g and b are only incident with one vertex. Such edges are called *loops*. The other edges are incident with two vertices. These two vertices are directly connected by such edges. Two directly connected vertices are *adjacent* or *neighbors*. The vertex z has no neighbors. Such a vertex is an *isolated* vertex. The edges d and h are both incident with u and x; such edges are *parallel*. The number of parallel edges between two vertices is the *multiplicity* of that direct connection between two vertices, also called the muliplicity of the edges.

Graven naar macht. Op zoek naar de kern van de Nederlandse economie, Amsterdam 1975, and Stokman, Roll Calls.

For a more elaborate introduction the reader is refered to Öre, Oystein, Graphs and Their Uses, New York 1963; Harary, Frank, Graph Theory, Reading/Mass. 1969, and Harary, Frank, et.al., Structural Models: An Introduction to the Theory of Directed Graphs, New York 1965.

[5] A directed graph is a graph in which each edge has a direction, from one vertex to another. An edge, together with its direction, is called an arc. In this paper we consider only undirected graphs. The stochastic scale model can also be applied in directed graphs; in that case the direction of the arcs should be taken into account.

A *simple graph* is a graph without loops and parallel edges. If parallel edges (can) occur in a graph, such a graph is called a *multigraph*. Information can be added to the edges and/or vertices of a graph. A graph, together with the information associated with its elements, is a *network*. Multigraphs can be represented as networks, if adjacent vertices are connected by only one edge and the multiplicity of the relation is added as information to that edge. In case of different kinds of relations between elements a multigraph can be used as representation by adding the kind of the relation as information to the edges. This situation will be considered in section 3. A graph is a *bipartite graph*, if the set of vertices of the graph is divided in two disjoint, non-empty subsets in such a way that the vertices have no neighbors within their own subset.

A graph or network can be used to *generate* other graphs or networks. Selection, aggregation and induction are the main processes used to generate a new graph or network from an existing one[6].

A new graph is generated by *selection*, if its elements (vertices, edges) are chosen from the set of elements of the original graph. The new graph then consists of the vertices and edges that satisfy the selection criteria. The selected edges must be edges of the new graph, the vertices incident with each selected edge must be selected vertices. A *subgraph* consists of a subset of the vertices of the original graph and all edges of the original graph of which both incident vertices belong to the selected subset of vertices. A subgraph is generated by deleting a number of vertices and all edges that are incident with these vertices. A *partial graph* consists of all vertices of the original graph and a subset of edges of that graph. A *partial subgraph* is a partial graph of a subgraph.

A new graph is generated by *aggregation of vertices*, if a subset of vertices in the original graph is condensed to a new vertex. Edges with both incident vertices within the subset became loops on the new vertex. Each edge between a vertex within and a vertex outside the subset becomes an edge between the new vertex and the vertex outside the subset. A new graph can also be generated by *aggregation of edges*, for example by condensing parallel edges to one new edge.

Induction is the third main process used to generate new graphs. In the case of induction two vertices in the new graph are directly connected by an edge, if these two vertices have a common neighbor in the original graph. The vertices of the induced graph (called inductees) are a subset of the vertices of the original graph. The induction can be limited to common neighbors in a certain subset of vertices of the original graph, called the subset of inductors. Each inductor induces edges between the inductees to which it is adjacent.

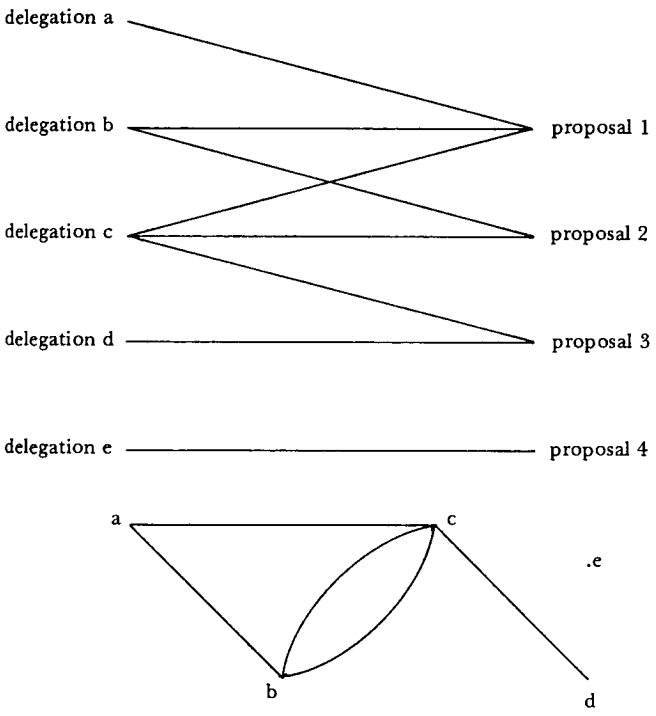
As an illustration, Figure 2 contains a bipartite graph. Each vertex in the first subset represents a delegation, each vertex in the other subset represents a proposal. A delegation and a proposal are connected by an edge if that delegation sponsored that proposal. This graph can be used to generate other graphs or networks, in particular those of co-sponsorship relations between delegations. The subset of in-

[6] Anthonisse, Jac M., and Lageweg, B. J., Graphlib O, Amsterdam 1975.

ductees are then the delegations, the subset of inductors the proposals. By induction a new graph is generated, consisting of relations between delegations. Each proposal induces an edge between each pair of delegations, by which it was sponsored. In the new graph two delegations are directly connected by as many edges as they co-sponsored proposals. The relations in the new graph are therefore the co-sponsorship relations between delegations. In this new graph parallel edges can be aggregated; the multiplicity of the edges (the number of co-sponsored proposals) is associated as information with each edge[7].

For the *analysis* of a graph or network we consider here only the density, the connectivity and the component density of a graph. We ignore the multiplicity of the edges. If the multiplicity is associated as information with the edges, we therefore ignore that information for the moment. Later in this section we shall take that information into account.

Figure 2: Induction



[7] The bipartite graph between delegations and resolutions could have been represented as a directed graph by associating a direction with each edge, e. g. from delegation to resolution. Analytically this has no meaning because only symmetric relations can be distinguished between delegations after induction, namely co-sponsorship relations.

The *density* (a) of a graph is the fraction of adjacent vertices. In a graph with p vertices and b neighbor relations (edges), the density is:

$$a = \frac{b}{1/2\, p(p-1)} \qquad (1)$$

In bipartite graphs vertices of the same subset cannot be adjacent. This can be taken into account by defining the density of a bipartite graph as:

$$a = \frac{b}{q \cdot r} \qquad (2)$$

in which q and r are the number of vertices in the two subsets.

If two vertices in a graph are not adjacent, they can be connected indirectly, *e. g.* because they have a common neighbor. If two vertices are adjacent, indirect connections between the two vertices can exist as well.

A *path* between vertex x and vertex y of a graph consists of an alternative sequence of edges l_i and vertices z_i:

$$x, l_1, z_1, l_2, z_2, \ldots, z_{k-1}, l_k, y \qquad (3)$$

in which l_1 is incident with x an z_1, l_2 incident with z_1 and z_2 etc. and l_k incident with z_{k-1} and y.

The vertices x and y are joined by a path through the vertices z_i and edges l_j.

A graph consists of one or more *components*. Two vertices belong to the same component if they are joined by a path; if two vertices are not joined by a path, they belong to different components. A graph is *connected* if it has only one component, *i. e.* if every pair of vertices is joined by a path. A component is therefore a maximal connected subgraph of the graph.

The connectivity of a graph is the number of pairs of vertices, joined by a path, as a fraction of all pairs of vertices in a graph. If the graph consists of s components and the i-th component consists of p_i vertices, the connectivity of the graph is:

$$c = \frac{2}{p\,(p-1)} \sum_{i=1}^{s} 1/2\, p_i\, (p_i - 1) \qquad (4)$$

The connectivity between two disjoint subsets Q and R of vertices of a graph is based on all pairs of vertices, of which one vertex belongs to one subset and the other to the other subset. If the i-th component contains q_i vertices from the first subset and r_i vertices from the other subset, the connectivity between the subsets Q and R is:

$$c_{QR} = \frac{1}{q \cdot r} \sum_{i=1}^{s} q_i\, r_i \qquad (5)$$

The density is based on all pairs of vertices that are directly connected; the connectivity is based on all pairs of vertices that are connected, either directly of indirectly.

Therefore we always have:

$$c \geqslant a$$

with equality if all pairs of vertices within all components are adjacent.

In the density we expressed the number of adjacent vertices as a fraction of all pairs of vertices. We can also express the number of adjacent vertices in a graph as a fraction of all pairs of vertices that belong to the same component, because vertices in different components cannot be adjacent. This *component density* is defined by:

$$a_c = \frac{b}{\sum_{i=1}^{s} 1/2 p_i\,(p_i - 1)} = \frac{2b}{\sum_{i=1}^{s} p_i\,(p_i - 1)} \qquad (6)$$

The relation between the connectivity, the component density and the density is:

$$a = a_c \cdot c$$

For bipartite graphs the component density is defined by:

$$a_c = \frac{b}{\sum_{i=1}^{s} q_i\, r_i} \qquad (7)$$

Again we have:

$$a' = a_c \cdot c_{QR}$$

One of the possibilities to take the multiplicity of the edges into account in the analyses is *a repeated (stepwise) analysis of the graph for different levels of multiplicity*. From the network we select a partial graph, containing only edges of a certain level of multiplicity or higher, and we analyze that partial graph; we repeat that for different levels of multiplicity. For example, we might analyze the network of co-sponsorship relations for the following levels of multiplicity (m):

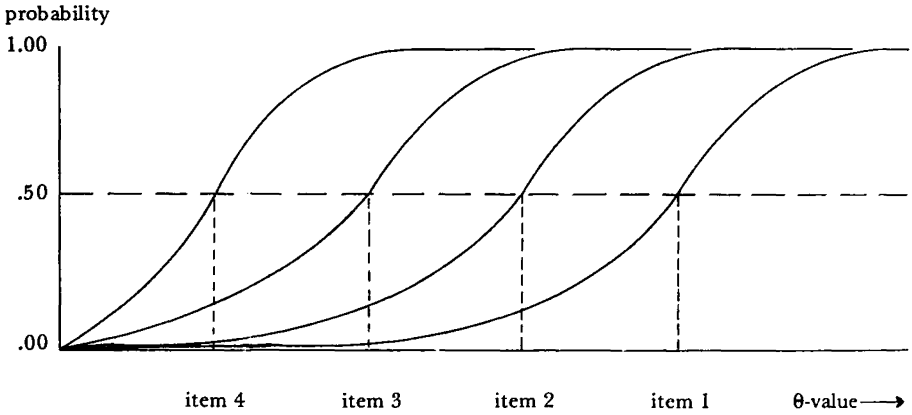
$$m \geqslant 10;\ m \geqslant 5;\ m \geqslant 2;\ m \geqslant 1$$

All edges of a partial graph at a higher level of multiplicity are contained in a graph at a lower level of multiplicity; the graph at a lower level of multiplicity is obtained by adding a number of edges to the partial graph at a higher level of multiplicity. The graph at a higher level of multiplicity is *nested* in the graph at a lower level.

2. Cumulative Scales of Vertices in a Simple Graph

In Figure 3 the stochastic cumulative scaling model of Mokken is given. We are dealing with an underlying continuum θ and two sets of elements (in the interview situation: subjects and items). Each subject has an unknown numerical value θ on the continuum. For each item we can draw a *trace line*, which gives the probability of a positive response for the different subject values θ. Each item also has an unknown numerical value, δ, on that continuum θ. For theoretical reasons we make the value δ_i for item i equal to the value θ of that subject that gives the positive response with probability of .5. In the stochastic scaling model of Mokken it is required that the trace lines are doubly monotone or holomorphic: each trace line is monotone non-decreasing for increasing subject values; the different trace line may not intersect (see Figure 3).

Figure 3. Trace Lines for Four Holomorphic Items



We can represent the data on which we apply the Mokken-model, as a bipartite graph. Each vertex in the first subset represents a subject, each vertex in the second subset represents an item. A subset and an item are directly connected by an edge if that subject responds positively to that item. The structure of the data is equal to that of Figure 2, the bipartite graph of sponsorship in the United Nations. It depends on the theoretical perspective whether we consider the delegations as the subset of subjects or as the subset of items. In Stokman and Stokman and Van Schuur the delegations were considered as items and the resolutions as subjects, because we were

interested in cumulative leadership scales of delegations[8]. We will now consider *which implications the cumulative scaling model has for the structure of the network between items as induced by the subjects.* In case of the sponsorship data we therefore consider which implications a cumulative structure of the relations in the bipartite graph between delegations and resolutions has for the network of co-sponsorship relations between delegations, which was generated from the bipartite graph by induction.

For a monotonely homogeneous set of k items it can be proven that all items, i, j (i, j = 1, 2, . . ., k) are positively correlated. In Table 2 two items are cross-tabulated, both in terms of frequencies and in terms of probabilities: n_i gives the frequency (π_i the fraction) of the subjects that respond positively to item i; n_j gives the frequency (π_j the fraction) of the subjects that respond positively to item j. n_{++} is the number of subjects (π_{++} the fraction) that respond positively to both item i and item j. For k items to form a cumulative scale it can now be proven that for all pairs of these k items:

$$\pi_{++} > \pi_i \cdot \pi_j \qquad \text{or}$$

$$\frac{n_{++}}{N} > \frac{n_i \cdot n_j}{N \cdot N}$$

Table 2. Cross-Tabulation of Two Items i and j

(a) in Terms of Frequency

		item j: +	item j: −	
item i	+	n_{++}	n_{+-}	n_i
	−	n_{-+}	n_{--}	$N-n_i$
		n_j	$N-n_j$	N

(b) in Terms of Probabilities

		item j: +	item j: −	
item i	+	π_{++}	π_{+-}	π_i
	−	π_{-+}	π_{--}	$1-\pi_i$
		π_j	$1-\pi_j$	1

The frequency n_{++} between items i and j is equal to the multiplicity of the edges between item i and j in the induced network. *Items form a cumulative scale only if the multiplicity of the edges between these items in the network induced by the subjects is larger than that expected in the case of random response*[9]. For the spon-

[8] See Stokman, Roll Calls and Stokman, Frans N., and Van Schuur, Wijbrand H., Leadership Dimensions among Developing Nations in the United Nations (reproduced in this volume).

[9] Mokken uses the coefficient Φ/Φ_{max} to measure the positive correlation between each pair of items. Loevinger's coefficient of scalability H is used to judge the scalability of a whole scale. This coefficient can be written as a function on the whole matrix of the Φ/Φ_{max} coefficients (H_{ij}'s) between each pair of items (Mokken, 1970, 150). The procecure of multiple scal-

sorship data it implies, that delegations form a cumulative (leadership) scale only if the multiplicity of the edges between these delegations in the network of co-sponsorship relations is larger than that expected in the case of random sponsorship.

Double monotony or *holomorphism* of the items can be checked by inspection of two matrices, one matrix containing the fractions π_{++} of all pairs and another matrix containing the fractions π_{--} of all pairs (see Table2 (b)). This check for holomorphism is described by Mokken. Let us consider in particular the matrix of the fractions π_{++}, because this matrix corresponds to the network between the items, as induced by the subjects. If the items are ordered according to their difficulty on the basis of the fraction of positive responses ($i < j : \pi_i \leqslant \pi_j$), the fractions π_{++} in each row of the matrix should increase monotonely with column index j. This is a necessary condition for a holomorphic set of items. This implies that the highest π_{++} fractions, corresponding with the highest multiplicities n_{++} in the network between items induced by the subjects, should be observed between the easiest items, *i. e.* the items with the largest number of positive responses. For the sponsorship data it implies, that the highest multiplicities n_{++} should be observed between the most active delegations. In Section 1 we considered a repeated analysis of a network. In this repeated analysis the minimal levels of multiplicities were decreased with each step. This stepwise procedure works out in a very specific way, if we are dealing with a network between items that form a doubly monotone cumulative scale. In each step the repeated analysis of the network then results in one dense component (and a number of isolated vertices). Moreover, over the consecutive steps with decreasing multiplicities items will be added to the component in the order of decreasing item difficulty: first, the two easiest item form one component; then, the next easiest items will be added to the component and so on until in the last step the most difficult item will be added to the component. *We may therefore conclude that the existence of one component with a high component density (a_c) at the different levels of multiplicity in an induced network indicates that the relations are cumulative in the graph that generated that network.*

ing selects items from a set of items in such a way that the scale coefficient H is maximized. It can be conceived as a cluster procedure on the network between items, as induced by the subjects. However, in this procedure we do not weight the edges according to their multiplicity (n_{++}): here we make the weight of an edge between items i and j equal to the correlation coefficient H_{ij}.

3. Cumulative Scales of Relations in a Multigraph

The stochastic scaling model of Mokken can also be used to test the cumulative character of relations in a multigraph[10]. Let us consider a network of p vertices and k different relations between these p vertices. We therefore consider k different graphs, defined on the same set of vertices. The k graphs can be represented in one multigraph by associating the relation as information with the edge. In the example of the co-sponsorship data we consider the co-sponsorship relations separately for k issues. We therefore consider k different graphs of co-sponsorship relations between the p delegations; two delegations are directly connected by an edge in the i^{th} graph, if they co-sponsored one or more proposals on issue i. The different *relations* can now be considered as the *items* in the scale model; each *pair of vertices* is considered as a respondent or *subject* in the scale model[11]; the *existence of an edge* in the i^{th} graph between a pair of vertices is equivalent to a *positive response* of a subject to that item in the scale model. The density in the i^{th} graph is related to the difficulty of the i^{th} item (the item difficulty) in the scale model. For a holomorphic set of items the order of the fractions π_i in the population (see Table 2) are determined by the order of the item difficulties δ_i. The sample fraction P_i is an unbiased and relatively precise estimate of the population fractions π_i[12]. The density of the i^{th} graph is the fraction of pairs of vertices between which the i^{th} relation has been observed. *The density of the i^{th} graph is therefore equal to the population fraction π_i or can be used as an unbiased and relatively precise estimate of it. The graph theoretical elaboration of the stochastic cumulative scaling model gives the density of the graph therefore a number of desirable conceptual and statistical properties.*

We conclude that the different cumulative dimensions of relations in a multigraph can be determined with the theory and procedures of scale analysis as developed by Mokken.

[10] The applicability of the theory and procedure of scale analysis in multigraphs was elaborated by Mokken and the author in the context of a research project on economic and political power.

[11] In table 2 N is therefore equal to 1/2 p (p−1), the total number of pairs of vertices in the graph.

[12] Mokken, Theory and Procedure, pp 126–8.

4. Generation and Analysis of the Networks for Each Cumulative Dimension of Relations

Scale analysis of the different relations in a multigraph, discussed in the last section, results in one or more cumulative scales of relations. These scales can now be used to generate networks between the vertices for each cumulative dimension separately. These networks can then be analyzed to determine the structure and central vertices for the different dimensions.

Suppose k relations formed a cumulative scale over pairs of vertices. In the example of the co-sponsorship data, suppose that co-sponsorship is cumulative over k issues. We can now determine the *summation score* for each pair of vertices over these k relations: the number of relations that has been observed between that pair of vertices. In the model of scale analysis this score is known as the score of a response pattern. It is a good estimate of the order of the pairs of vertices (the subjects) on the underlying dimension of relations[13]. *For each cumulative dimension of relations we therefore generate a network in which the weight of the edges is equal to the summation score over the k relations in the scale.* These networks can then be analyzed with graph theoretical concepts and measures to determine the structure and central vertices.

An example of such an elementary graph analysis of networks for different cumulative dimensions of relations can be found in Stokman[14]. In the period 1960–63 two cumulative dimensions of co-sponsorship relations existed among the developing nations: one of colonial issues and one of socio-economic issues. We therefore generated two networks among the developing nations: a colonial network and a socio-economic network, the weights of the edges being the summation score over the issues in the issue dimension. Particularly the relations within Latin America and those between Latin America and Afro-Asia were quite different for the two networks.

[13] Op. cit., pp. 128–129.
[14] Stokman, Roll Calls.